Evaluating Inter-Annotator Agreement for Non-Manual Markers in Sign Languages

Lyke D. Esselink, Marloes Oomen, Floris Roelofsen
University of Amsterdam
Amsterdam, the Netherlands
{l.d.esselink, m.oomen2, f.roelofsen}@uva.nl

Abstract

This paper is part of a larger project that aims to create a standardized procedure for annotating non-manual markers (NMMs) in sign language data. The paper describes two approaches to evaluating inter-annotator agreement, the *event-based* approach and the *frame-based* approach, and uses a combination of these two approaches to evaluate the annotation guidelines introduced in Oomen et al. (2023). The evaluation reveals that for several labels in the annotation scheme inter-annotator agreement is rather low. This indicates that the annotations guidelines need to be further improved. We present concrete recommendations for how this may be achieved, and intend to implement these recommendations in future work. All data and analysis scripts are available.

Keywords: sign language, non-manual markers, annotation guidelines, inter-annotator agreement

1. Introduction

This paper is part of a larger project that aims to create a standardized procedure for annotating non-manual markers (NMMs) in sign language data. The initial steps we took as part of this project—developing annotation guidelines and creating a dataset annotated according to these guidelines by two annotators—were previously reported in Oomen et al. (2023). In the present paper, we report on the next step: a thorough evaluation of inter-annotator agreement, yielding substantial recommendations for improvement of the guidelines.

In Section 2, we outline our general motivations for developing a new protocol for annotating NMMs. Section 3 provides a brief summary of the first steps towards such a protocol as reported in Oomen et al. (2023). In Section 4, we describe two general methods for evaluating inter-annotator agreement which can be applied to sign language data. Section 5 discusses the results of applying these methods to our test dataset, leading to several recommendations for further improving our annotation guidelines. This is the main contribution of the paper. Sections 6 discusses some methodological prospects and limitations of the evaluation methods we adopted, and Section 7 concludes. Before the bibliography, we provide pointers to all supplementary materials: the annotation guidelines, evaluation data, analysis scripts, and a technical report with extensive discussion of all results.

2. Motivation for the Larger Project

In sign languages, facial expressions, body movements, and other NMMs serve a wide range of linguistic functions, in addition to the gestural and affective functions they may fulfil more generally.[1] There are plenty of examples in the literature tying particular NMMs (or clusters of NMMs) to particular grammatical functions (for a recent overview, see Wilbur, 2021). For instance, Bahan (1996) has argued that eye gaze (or head tilt in the case of first person) can be used to mark verb agreement in American Sign Language (ASL); Göksel and Kelepir (2013) have claimed that (forward or backward) head tilt in Turkish Sign Language marks interrogative mood while specific combinations of head tilt and head movement distinguish polar (forward + head nod) and content (backward + headshake) questions; Wilbur and Patschke (1998) have proposed, again for ASL, that body leans are used to convey contrast at the prosodic, lexical, semantic, and pragmatic level. Works such as these provide highly valuable descriptive, analytical and theoretical insights, but they tend to be based on relatively small sets of examples, for which it is often unclear exactly how they were obtained or analyzed. The analyses also generally do not involve detailed qualitative annotation of NMMs, or the annotation procedure is not discussed.[2] Moreover, (individual) variation in NMMs use is often not considered. This means that many claims about NMMs and their properties and functions in sign languages still await robust empirical verification, which cannot be done without in-depth analysis of NMM patterns by means of careful annotation of linguistic data.

Facial expressions and other NMMs also play

[1] This section overlaps to a large extent with Section 2 from Oomen et al. (2023).

[2] Notable exceptions include Pendzich (2020) on lexical NMMs in German Sign Language and Lackner (2017) on the various functions of head and body movements in Austrian Sign Language.

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 66–76

an important role in multimodal communication, where they have been shown to be connected to a wide variety of semantic, pragmatic, and social functions (e.g., Bavelas and Chovil, 2018; González-Fuente et al., 2015; Nota et al., 2021; Tomasello et al., 2019). Thus, research in this domain likewise requires (and sometimes already includes; e.g., González-Fuente et al. 2015, Nota et al. 2021) fine-grained annotation of facial expressions and other visual cues in video data.

Annotation of NMMs is highly time-consuming and also poses challenges for data analysis, given the considerable number of possible NMMs and the fact that temporal information is ideally also taken into account. Even so, as we have discussed, such work is vital both for empirical assessment of theoretical claims as well as to gain more insight into the factors that lead to variation in NMMs use in sign language and multimodal communication.

Currently, the field lacks standard guidelines for annotating NMMs. That is to say, guidelines for annotating NMMs do exist, but none have been thoroughly validated and have become a community-wide standard. Researchers studying NMMs often end up devising new annotation protocols tailored to their specific research objectives.[3] Furthermore, we also lack a standard method to quantify inter-annotator agreement. In fact, publications in sign language linguistics rarely report inter-rater agreement scores. For instance, ten out of the seventeen research articles published in *Sign Language & Linguistics* in 2021-2023 investigate properties of sign languages based on annotated video data, but just one of them reports inter-annotator agreement scores. Adopting a standard method for this purpose would benefit the field by increasing data transparency, and would enable us to iteratively evaluate and improve our annotation guidelines.

The general project that the present paper is part of therefore pursues (i) the development of a reliable protocol for the annotation of NMMs, and (ii) a procedure for evaluating inter-annotator agreement. This paper focuses on the second project pillar. Indeed, it does not really matter for the purpose of this paper which annotation protocol we evaluate.

[3] A reviewer made us aware of an extensive annotation protocol for both manual and non-manual markers that was developed in the context of the SignStream project (Neidle, 2002). While this annotation scheme has to our knowledge not been evaluated for inter-annotator agreement, some of the general and specific insights and recommendations discussed in these guidelines overlap with those discussed in the present paper. We thank the reviewer for pointing us to this work, and we will briefly return to it in our discussion on the distinction between *poses* and *movements* in Section 5.1.

3. Summary of Oomen et al. (2023)

In Oomen et al. (2023) we presented a first version of the annotation guidelines, according to which two coders annotated a test set of 60 interrogative sentences in Sign Language of the Netherlands (NGT), which came from a larger dataset created in the context of another study. The annotations were produced in ELAN (2023). Coder 1 (C1) annotated 585 events over the 12 tiers specified in the guidelines, and Coder 2 (C2) annotated 564 events. The tiers concerned the eyebrows, eye shape, eye gaze direction, shoulder position, body position, head position and movement, mouth configuration, lip corner configuration, and nose wrinkle.

In Oomen et al. (2023), we already briefly evaluated the reliability of the resulting annotations and included a few recommendations for the improvement of the annotation guidelines. However, the discussion was limited to one annotation tier (concerning the eyebrows) and one evaluation method. In the present paper, we provide a more in-depth evaluation, and offer more extensive recommendations to improve the guidelines.

4. Evaluation Methods and Measures

Video-recorded sign language data represents so-called *timed-event sequential data* (Bakeman et al., 2009; Bakeman and Quera, 2011). In general, such data involve recordings of sequences of events, each with a particular time duration. Besides sign linguists, researchers investigating other phenomena (e.g., speech, multimodal communication, or animal behavior) also work with this kind of data, make similar use of annotations, and have devised several methods to assess inter-annotator agreement for this type of data. Broadly, two approaches can be distinguished: *frame*-based approaches and *event*-based approaches (Bakeman et al., 2009).[4] In both these approaches, inter-annotator agreement is quantified using *confusion matrices* and *agreement indices*. We briefly explain each of these methods in this section.

4.1. The Event-Based Approach

In the event-based approach, we treat all annotations as 'events', and first determine the temporal overlap between annotations of the two coders, who we refer to as C1 and C2. This is done separately for each tier. For this approach, the annotation label 'neutral' (used when a particular facial feature or body part is in a neutral position) is not classified as an event, so these labels are disregarded. Two annotations are taken to 'match' if their

[4] Frame-based approaches are also referred to as *time*-based approaches (Bakeman et al., 2009).

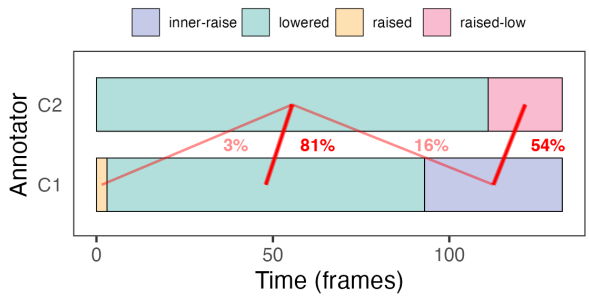

Figure 1: Annotations for the eyebrow tier of a sentence. Red lines show the percentage overlap between all annotations; the thick lines show the percentage overlap between 'matching' annotations.

overlap exceeds a pre-defined overlap threshold. At this stage, the label *values* are not considered: matches are established purely based on temporal overlap. We use an overlap threshold of 51%. Overlap between two annotations i and j is calculated according to the following formula (Holle and Rein, 2015):

$$O_{ij} := \frac{\min(\mathit{offset}_i, \mathit{offset}_j) - \max(\mathit{onset}_i, \mathit{onset}_j)}{\max((\mathit{offset}_i - \mathit{onset}_i), (\mathit{offset}_j - \mathit{onset}_j))}$$

In words, O_{ij} is the length of the overlap between i and j divided by the length of the longest of the two annotations. If O_{ij} does not exceed the threshold, i and j are not regarded as a match. If an annotation by C1 does not have any matching annotations by C2, that annotation is regarded as 'unmatched'.

Figure 1 shows the annotations by C1 and C2 for the eyebrow tier of an example sentence in our test dataset. The red lines show the percentage overlap between all annotations of C1 and C2, respectively. The thin transparent lines show the percentage overlap between 'unmatched' annotations, while the 'matching' annotations are illustrated by the thick opaque lines. Again, note that 'matching' annotations do not necessarily involve the same label, the only criterion is that they have sufficient temporal overlap. We turn to quantifying the extent to which matching annotations agree in terms of their labels in Section 4.3.

4.2. The Frame-Based Approach

On the frame-based approach, we simply consider each individual frame in all videos annotated by C1 and C2, and then determine whether the labels applied by C1 and C2 to each of these frames correspond. We do this separately for each tier. On this approach we do take 'neutral' labels into account, so that for each frame we can compare the labels that the two coders assigned.

4.3. Confusion Matrices

Both on the event-based approach and on the frame-based approach, the first step in quantifying inter-annotator agreement is to compile a so-called *confusion matrix*. For examples of confusion matrices for two of the tiers we evaluated, see Section 5.2 and 5.3. Cell ij in a confusion matrix displays the number or the percentage of events/frames which C1 labeled as i and C2 labeled as j. When displaying percentages, a confusion matrix is either constructed from the perspective of C1 (which means that all rows add up to 100%) or from the perspective of C2 (all columns add up to 100%).

4.4. Agreement Indices

Besides confusion matrices, another way to quantify inter-annotator agreement is to compute *agreement indices* for each label. Here it is important to note that so-called *raw* agreement indices are insufficient. To illustrate this, suppose that two annotators x and y label 100 items. To 50 items they both apply label A, to 20 items only x applies label A, to 20 items only y applies label A, and to the final 10 items they both apply another label. Then, x and y agree in 50 + 10 = 60 of the cases as to whether label A applies or not. The raw agreement index for label A, then, is 0.6. However, this does not take into account the possibility that, at least in some cases, x and y may have agreed on the application of label A *by mere chance*. Both x and y applied label A to 70% of the items, and other labels to 30% of the items. If they would randomly assign label A to 70% and other labels to 30% of the items, they would agree 58% of the time as to whether A applies or not (because $(0.7 * 0.7) + (0.3 * 0.3) = 0.58$). So the raw agreement index, $i_{raw} = 0.6$, is just slightly higher in this case than the chance agreement index, $i_{chance} = 0.58$. Chance-corrected agreement indices take this factor into account.

One widely used chance-corrected index is Cohen's κ (Cohen, 1960). It is computed by dividing the difference between i_{raw} and i_{chance} by the difference between i_{chance} and the index for perfect agreement, which is 1.

$$\kappa := (i_{raw} - i_{chance})/(1 - i_{chance})$$

In the example above, κ would amount to $0.02/0.42 = 0.05$. To give some other examples, if $i_{raw} = 0.7$ and $i_{chance} = 0.5$ then $\kappa = 0.4$, and if $i_{raw} = 0.9$ and $i_{chance} = 0.6$ then $\kappa = 0.75$.

It is important to note that it is not straightforward to interpret agreement indices such as Cohen's κ. Some researchers have proposed specific interpretations. For instance, a frequently cited interpretation is that of Landis and Koch (1977, 165), who posit that a κ score of 0.21–0.40 amounts to 'fair' agreement, 0.41–0.60 to 'moderate' agreement, 0.61–0.80 to 'substantial' agreement, and

0.81–1 to 'almost perfect' agreement. However, it has been noted in the literature that such absolute interpretations are arbitrary and problematic, because κ scores can be affected by *label prevalence* (whether the labels are equiprobable or not), *coder bias* (whether the marginal probabilities for the two coders are similar or different), and the *number of possible labels* for a given annotation tier (Bakeman et al., 1997; Sim and Wright, 2005).

Thus, not too much should be read into any single κ score on its own. Rather, a κ score should always be considered *relative to other* κ *scores*. For instance, if there are three roughly equiprobable labels for a given annotation tier (A, B, C), and the κ score for A is much lower than that for B and C, then we can conclude that the instructions for label A in the annotation guidelines were less reliable than those for B and C. Another possibility is to compare κ scores across iterations of the annotation guidelines. With every new iteration, we hope to obtain higher κ scores. If we do, this confirms that the adjustments we made indeed succeeded in making the protocal more reliable. The latter type of comparison is our main intended use of κ scores. That is, we mainly report κ scores here for comparison with future iterations of the guidelines.[5]

5. Results and Recommendations

We have compiled confusion matrices and κ scores for all twelve tiers in the annotation guidelines, based on the test dataset from Oomen et al. (2023) described above, both under the event-based approach and under the frame-based approach. Based on our analysis and comparison of these twelve tiers, we formulate a number of general recommendations for improvement of the annotation guidelines in Section 5.1. For reasons of space, we cannot discuss the results for all tiers individually; they are presented in a technical report which is available in the supplementary materials. Here, we only discuss two specific tiers, *head y* (with labels 'up', 'down' and 'neutral'; Section 5.2) and *head move* (with labels 'nod', 'nodding', 'shake', 'shaking', 'sideways', and 'neutral'; Section 5.3), as they relate to many issues that we target with our general recommendations.

5.1. General Recommendations

The most important general insight we obtained is that a methodical distinction should be made between two types of NMM, which we refer to as *poses* and *movements*. As a reviewer pointed out, a similar distinction is made in the SignStream annotation protocol (Neidle, 2002), namely a distinction between 'positions' and 'movements'. The former involve some part of the face or body 'first moving to a target position and then maintaining that position' for some time, while the latter involve 'continuous (potentially repeated) movements' (Neidle, 2002, p.24).

Very much in line with this, we define a *pose* as a non-manual feature which can be characterized in terms of a *single configuration* of part of the face or body, which is *held* for a certain amount of time. Disregarding transitional movements in and out of a pose (see below for discussion on how to treat such transitions), a pose itself does not involve inherent movement. Clear examples of poses are the features 'head up' and 'head down' on the *head y* tier (see Section 5.2). Poses can in principle be labeled on a frame-by-frame basis.

On the other hand, we define *movements* as non-manual features for which a *temporal progression* from a certain starting configuration, possibly through certain intermediate configurations, to a certain target configuration is characteristic. Movements typically happen within a relatively short amount of time. Many movements are oscillatory; in this case the target configuration is the same as the starting configuration. Clear examples of movement NMMs are head nods and headshakes on the *head move* tier (see Section 5.3), and eye blinks. Since movements cannot be characterized in terms of a single configuration but involve a temporal progression through multiple configurations, they can never be identified based on a single video frame only. Labeling a video segment as involving a certain movement is thus qualitatively different from labeling it as involving a certain pose, as the entire sequence of frames within the given segment—and not each frame individually—determines the annotation value.[6]

This discussion yields three concrete recommendations that should be integrated in future versions of the annotation guidelines.

Firstly, in the current version of the annotation guidelines, certain tiers contain labels for both *poses* and *movements*, as exemplified by the *head move* tier discussed in Section 5.3. Given the

[5] The event-based method of Holle and Rein (2015) that we have described in this section is implemented in ELAN and can be performed straightforwardly by selecting File → Multiple File Processing → Calculate Inter-Annotator Reliability. The output is a .txt file with agreement matrices and Cohen's κ. We have re-implemented the method in R with additional visualisation functionalities (see Section 9 for a link to the documented R script). Advantages of the R script over the ELAN functionality are (i) that it is fully transparent and (ii) that it can easily be modified and extended (see Section 6 for some suggestions in this direction), and (iii) that the results can be visualised in various ways.

[6] An analogy: movement labels are like *collective predicates*, while pose labels are like *distributive predicates*.

qualitative differences between poses and movements that we just identified, annotation tiers should comprise either poses or movements, not both. It should also be made explicit for each annotation label whether it describes a pose or a movement. This is lacking in the current guidelines, and it is evident that this sometimes led to confusion among coders. For instance, the label 'closed' on the *eye gaze* tier was applied differently by our coders. One coder used it only to label longer segments where the signer kept their eyes closed (an eye *pose*). The other coder used the label in such cases too, but also applied it to short eye blinks (an eye *movement*). Section 5.2 discusses another example.

Secondly, on *pose* tiers, both neutral and non-neutral configurations (e.g. 'head neutral' vs. 'head up' or 'head down') should be annotated, because neutral configurations are poses as well. As a consequence, *pose* tiers are typically *continuous*, in the sense that every video segment is given some label.[7,8] In contrast, on *movement* tiers, only movement events should be annotated; if there is no movement that corresponds to one of the labels on the tier, nothing should be annotated. For example, on a tier for eye blinks, each blink should be labeled, but no further annotations should be added; 'neutral' is not a useful label in this case since it does not describe a movement.

Finally, the guidelines should specify what it means for a *pose* to be held "for a certain amount of time", and for a *movement* to occur "within a relatively short amount of time". For instance, if we specify within which time frame a signer's eyes should close and re-open for it to be considered an *movement*, i.e. a blink, instead of a *pose*, then coders can make a principled distinction between these two labels in situations where there may otherwise be confusion. We plan to undertake empirical work to determine suitable thresholds.

Relatedly, there is the issue of when a *pose* or *movement* should start and end. This issue is particularly tricky when it comes to *poses*: at what point should a coder decide that a signer's eyebrows are no longer in, say, a 'neutral' position, but have rather become 'raised'? As a basic principle, we propose that pose annotations should include the transition movement *into* the pose but not the one *out of* that pose (and into the next one).[9]

[7]There are exceptions to this. For instance, on the pose tier for eye gaze direction, segments in which the eyes are closed need not be given a label.

[8]What we suggest here for poses differs from the treatment of 'positions' in the SignStream protocol; neutral positions are not regarded there as true positions and as such are not annotated.

[9]This differs, again, from the SignStream protocal, where transition movements in and out of positions are coded separately, as 's(tart)' and 'e(nd)', respectively.

Another important insight we obtained concerns tier structure. With twelve tiers, the current guidelines already contain a fairly elaborate tier structure, yet we found that further distinctions between tiers and/or annotation labels are desired for reasons of clarity, exhaustiveness, and systematicity. Moreover, an extensive tier structure makes it easier for researchers to focus on only specific NMMs. We therefore propose the following principles for systematic expansion of the tier structure: (1) Every tier should concern a UNIQUE BODY PART (e.g. head, eyelids, nose, eyebrows); (2) Every tier should only include labels for *poses*, or only for *movements* (e.g. the eyelid *movement* 'blink' should be annotated on a different tier than the eyelid *pose* 'closed'); (3) Every tier should contain labels that are MUTUALLY EXCLUSIVE (i.e., any two NMMs that can co-occur should be annotated on separate tiers); (4) The set of labels for *pose* tiers should be JOINTLY EXHAUSTIVE – i.e., each *pose* tier should have a set of labels that cover the full range of possible *poses* for the relevant body part (as discussed above, this does not apply to *movement* tiers; (5) The set of labels on a given tier should be sufficiently CONTRASTIVE.

Regarding criterion (5), some tiers in the current guidelines include pairs of labels that describe the same NMM but to different degrees of engagement (e.g., 'squint-full' and 'squint-half' on the 'eye shape' tier). Our analyses show that the inclusion of such labels generally lead to poor inter-coder agreement. We suggest to only include the label 'squint' in future versions of the guidelines.

While it seems impossible to reliably annotate the degree of engagement of non-manual features, we do believe it is useful to obtain a measure of the *confidence level* of the coders (previously explored, for instance, for annotation of emotions in text by Troiano et al. 2021). Coders may record, for every annotation event, their level of confidence in the label they applied, on a three-point scale from low to high. Researchers then have the option to only analyze a subset of the data with high confidence scores, and to compare this analysis to one taking the entire dataset into account. Moreover, confidence ratings would be useful as training data for machine learning in the future.

In such a system, including 'neutral' poses in the repertoire of possible poses is important. Say a study only wishes to include annotations with high confidence ratings, but 'neutral' poses are not labeled to begin with. Then for all events that are not considered, it is unknown whether they are not included because they received a low confidence rating or because they involve a neutral state.

Besides a sub-tier for confidence ratings, another sub-tier we propose to add is one on which annotators can indicate when a particular non-manual feature clearly does not have a communicative func-

tion, e.g. when a signer wrinkles their nose because it's itching, or turns their head because of an unexpected movement next to them. In such cases, coders can make a note on this tier, allowing for irrelevant events to be excluded from the analysis.

Furthermore, poses and especially movements should be illustrated in the guidelines not just with static video stills but also with video clips or GIFs. As such, the next version of the guidelines should be constructed in digital format such as in the form of a website or a slide deck.

A final recommendation does not concern the guidelines, but rather the data collection method. A major challenge that arises when manually annotating video data is that it involves analyzing 2D data that represents a 3D reality. Specifically, we found that a single (near-)frontal camera view makes the work for manual coders particularly challenging. We therefore advise researchers collecting data to always use multiple cameras, including a side-view camera. In addition, 3D capturing techniques may be considered as well (see Esselink et al., 2023).

5.2. *Head y*

On the *head y* tier (a *pose* tier) there were three possible labels: 'up', 'down', and 'neutral'.

Frame-Based Approach The confusion matrices in Table 1 show that the coders generally agreed on the 'neutral' label, but not on 'down' and 'up'.

Event-Based Approach The event-based confusion matrices in Table 2 show that the two coders identified a similar number of events as 'down' or 'up' events. However, the agreement rates concerning these events are extremely low. In total, only 15% of the 68 events annotated on this tier matched another event with the same label.

Error Analysis To better understand the low agreement scores for this tier, we carried out an error analysis of the mismatched events. We found that 3/19 [3/23] unmatched events labeled as 'down' by C1 [C2] were unmatched due to the coders not agreeing on onset and/or offset, resulting in insufficient overlap between the events to establish a match. For 2/19 [4/23] events, C1 [C2] had labeled (almost) the entire sentence as 'down', but C2 [C1] labeled two short events as 'down', which were preceded and followed by 'neutral' interludes. For the remaining 14/19 [16/23] unmatched events, C1 [C2] had identified (usually quite short) parts of the sentence as 'down' events, whereas C2 [C1] labeled these segments as 'neutral'.

For all unmatched 'up' events, one of the coders labeled the relevant segment as 'neutral'.

Cohen's Kappa On the frame-based approach, the κ scores are very low: 0.27 ('down'), 0.27 ('up'), and 0.21 ('neutral'). On the event-based approach, they are even worse: -0.27 ('down') and 0.14 ('up').

Tier-specific Recommendations The results for the *head y* tier show that the coders were hardly consistent with each other in identifying 'up' and 'down' events. In most cases, the disagreements were *categorical*, i.e., one coder identified an 'up' or 'down' event while the other coder labeled the same segment as 'neutral'.

Based on these results, we have three specific recommendations for this tier. First, we expect that use of a second camera offering a side view would facilitate more accurate and consistent coding of head position. Second, the annotation guidelines need to be more explicit on how much the head should diverge from a neutral position in order for it to count as a head 'up' or 'down' event. And third, the guidelines should specify a minimum duration of 'up' and 'down' events, in particular so as to distinguish 'down' events from *head nods* (see Section 5.3 below). In future work, we aim to establish concrete minimum duration values to be included in the guidelines.

5.3. *Head move*

The *head move* tier is intended for annotating head *movements*, and includes the labels 'nod' (single nod), 'nodding' (multiple nods), 'shake' (single shake), 'shaking' (multiple shakes), 'sideways' (single sideways movement of the head), and 'neutral'.

Frame-Based Approach For this tier, there is generally not much confusion between the coders. One might have expected low agreement on the labels 'nod' vs 'nodding', and 'shake' vs 'shaking', but Table 3 shows that this is not necessarily the case. However, we can make some other interesting observations pertaining to these labels.

Overall, C2 applied the various labels (other than 'neutral') to more frames than C1, who used 'neutral' more often. An especially interesting pattern can be observed for the label 'nod': when C1 applied this label, C2 agreed 52% of the time, labeling the remaining frames as 'nodding' (23%) or 'neutral' (25%). When C2 used 'nod', C1 only agreed 26% of the time. The remaining 74% of frames were labeled overwhelmingly as 'neutral' (69%). Both coders applied 'nodding' quite similarly, although C2 again labeled more frames as such than C1.

For 'shake' and 'shaking', we see a large disparity in application for both coders. The label 'shake' is barely assigned to any frames, totalling only 87 frames for C1, and 57 frames for C2. In contrast, the label 'shaking' is applied to a large number

Table 1: Confusion matrix for the *head y* tier showing the total number of frames (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of frames

C1/C2	down	up	neutral	Total
down	597	24	1086	1707
up	6	102	165	273
neutral	720	273	4920	5913
Total	1323	399	6171	7893

(b) C1

C1/C2	do	up	ne	Total
do	**35**	1	64	100
up	2	**37**	61	100
ne	12	5	**83**	100

(c) C2

C1/C2	do	up	ne
do	**45**	6	17
up	0	**26**	3
ne	55	68	**80**
Total	100	100	100

Table 2: Confusion matrix for the *head y* tier showing the total number of events (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of events

C1/C2	down	up	unmatched	Total
down	7	0	19	26
up	0	3	6	9
unmatched	23	10	0	33
Total	30	13	25	68

(b) C1

C1/C2	do	up	un	Total
do	**27**	0	73	100
up	0	**33**	67	100
un	70	30	0	100

(c) C2

C1/C2	do	up	un
do	**23**	0	76
up	0	**23**	24
un	77	77	0
Total	100	100	100

of frames, totalling 1704 frames for C1, and 1926 for C2. Again, we see a similar pattern as above, where C2 assigned this label to more frames than C1, who mostly labeled these remaining frames as 'neutral'. However, in this case there is a higher level of agreement: C2 agreed with the 'shaking' labels applied by C1 99% of the time, and C1 agreed with C2 88% of the time.

Finally, C2 applied the label 'sideways' to 144 frames, which were all labeled as 'neutral' by C1. C1 never applied the label 'sideways'.

Event-Based Approach The confusion matrices in Table 4 for the event-based approach show the same general patterns as the confusion matrices of the frame-based approach in Table 3. There is barely any confusion between the labels 'nod'/'nodding' and no confusion between the labels 'shake'/'shaking'. Looking closer at the data, we see that the confusion between these labels for the frame-based approach can be mostly attributed to disagreement on the onsets and offsets of events.

The labels 'nodding', 'shake', and 'shaking' were applied to a similar number of events by both coders, with the total number of events assigned one of these labels differing by only 1. This shows that, as C2 generally applied these label to more frames than C1, the annotation events by C2 were likely longer in duration than those of C1. For the label 'nod', we see a big disparity in the number of annotation events: C1 labeled 15 events as such, while C2 assigned this label to 26 events. The majority of these events were unmatched for both C1 and C2. The labels 'shake' and 'sideways' were barely assigned to any events by the coders.

Error Analysis A possible explanation for the disparity between the frames and events labeled as 'nod' by C2 and as 'neutral' by C1 is that C1 labeled these instances as 'down' (in the *head y* tier) instead. We briefly examine this possibility here; Table 5 shows the events of interest. In 19 cases, C2 labeled an event on the *head move* tier as 'nod' while C1 labels it as 'neutral'. We examine labels given to corresponding events in the *head y* tier. The rows display the labels given to these events by C1; the columns display the labels given to the matching event by C2.

In 9 cases, C1 labels a corresponding event on the *head y* tier as 'down'; of these 9 cases, C2 labels the corresponding event as 'down' twice, and as 'neutral' 7 times. However, also in 9 cases, C1 labels a corresponding event on the *head y* tier as 'neutral'; of these, C2 labels the corresponding event as 'down' 3 times, and as 'neutral' 6 times. In one case, C1 labels the corresponding event as 'up', while C2 labels this event as 'neutral'.

Therefore, we see that in about 50% of the cases examined here, C1 labeled the events as 'down' on the *head y* tier instead of 'nod' on the *head move* tier. We cannot definitively conclude that in these cases, C1 labeled events as 'down' in the *head y* tier in lieu of labeling the corresponding events as 'nod' in the *head move* tier. However, this does explain some of the discrepancy.

This leads us to another interesting observation. C2 labeled 5 events as 'nod' in the *head move* tier, as well as labeling a simultaneous event as 'down' in the *head y* tier. We find that for 4 of these occurrences, the events on both tiers have roughly the same onsets and offsets. A quick check of the annotations provided by C1 reveals 4 'nod' events

Table 3: Confusion matrix for the *head move* tier showing the total number of frames (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of frames

C1/C2	nod	nodding	shake	shaking	sideways	neutral	Total
nod	183	81	0	0	0	90	354
nodding	27	567	0	3	0	60	657
shake	0	0	51	21	0	15	87
shaking	6	0	0	1686	0	12	1704
sideways	0	0	0	0	0	0	0
neutral	489	240	6	216	144	3996	5091
Total	705	888	57	1926	144	4173	7893

(b) C1

C1/C2	nd	ng	se	sg	si	ne	Total
nd	**52**	23	0	0	0	25	100
ng	4	**86**	0	0	0	9	100
se	0	0	**59**	24	0	17	100
sg	0	0	0	**99**	0	1	100
si	0	0	0	0	**0**	0	0
ne	10	5	0	4	3	**78**	100

(c) C2

C1/C2	nd	ng	se	sg	si	ne
nd	**26**	9	0	0	0	2
ng	4	**64**	0	0	0	1
se	0	0	**89**	1	0	0
sg	1	0	0	**88**	0	0
si	0	0	0	0	**0**	0
ne	69	27	11	11	100	**96**
Total	100	100	100	100	100	100

Table 4: Confusion matrix for the *head move* tier showing the total number of events (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of events

C1/C2	nod	nodding	shake	shaking	sideways	unmatched	Total
nod	6	0	0	0	0	9	15
nodding	1	9	0	0	0	4	14
shake	0	0	3	0	0	1	4
shaking	0	0	0	19	0	3	22
sideways	0	0	0	0	0	0	0
unmatched	19	4	0	4	5	0	32
Total	26	13	3	23	5	17	87

(b) C1

C1/C2	nd	ng	se	sg	si	un	Total
nd	**40**	0	0	0	0	60	100
ng	7	**64**	0	0	0	29	100
se	0	0	**75**	0	0	25	100
sg	0	0	0	**86**	0	14	100
si	0	0	0	0	**0**	0	0
un	60	12	0	12	16	**0**	100

(c) C2

C1/C2	nd	ng	se	sg	si	un
nd	**23**	0	0	0	0	53
ng	4	**69**	0	0	0	24
se	0	0	**100**	0	0	6
sg	0	0	0	**83**	0	18
si	0	0	0	0	**0**	0
un	73	31	0	17	100	**0**
Total	100	100	100	100	100	100

in the *head move* tier with simultaneous events in the *head y* tier labeled as 'down' or 'up'. However, the onset and offset of events in these tiers do not match up, meaning that the head was angled as either 'down' or 'up' for a longer period of time, within which a 'nod' took place. We can conclude that C1 did not confuse the meaning of 'nod' on the *head move* tier and 'down' on the *head y* tier, whereas the difference between these labels was not always clear for C2.

Cohen's Kappa For the frame-based approach, the κ indices for 'nodding' (0.71), 'shake' (0.71), and 'shaking' (0.91) are reasonably high, as expected.

C1/C2	down	neutral	up	Total
down	2	**7**	0	9
neutral	3	6	0	9
up	0	1	0	1
Total	5	14	0	19

Table 5: Labels given to events in the *head y* tier, occurring simultaneously with events in the *head move* tier, which have been labeled as 'neutral' by C1, and 'nod' by C2

The κ index for 'nod' (0.30) is much lower, as there was a lot of disagreement about this label between the coders. The κ index for sideways is 0.00, as the coders never agreed on this label.

The κ indices for labels in the event-based approach are generally lower than those of the frame-based approach. However, the indices are still relatively high for 'nodding' (0.61), 'shake' (0.85), and 'shaking' (0.79). The index for 'nod' is lowered to 0.09, while the index for 'sideways' remains 0.00.

Tier-specific Recommendations Firstly, we note that all labels on the *head move* tier can be categorized as (oscillating) *movements*, with the exception of 'sideways', which is a *pose*. The latter should therefore be moved to a separate *pose* tier.

Secondly, although head nods and headshakes involve the same body part, are mutually exclusive, and contrastive (see Section 5.1), we recommend that head nods and headshakes are annotated on separate tiers because they serve very different functions in sign languages. This way, researchers interested only in headshakes need not annotate head nods and vice versa.

Finally, the annotation guidelines should include clear descriptions of what constitutes a 'nod' (*movement*) and a head 'down' (*pose*), with concrete temporal indications for the required length of *movements* vs. *poses* (in terms of time rather than frames, as users may use different frame-rates). The guidelines should warn that these features can look similar, and show examples of the differences between them.

6. Discussion of Evaluation Methods

Considering the assessment of inter-annotator agreement for timed-event sequential data in general, Bakeman et al. (2009, 146) advise the use of both event-based and frame-based methods, as "each provides somewhat different . . . but valuable information as to how observers are disagreeing, and are thus useful in different ways as observers strive to improve their agreement". We will now briefly discuss some concrete benefits of these methods we identified for NMM data.

An advantage of the event-based approach is that it allows for an error analysis, as illustrated in Section 5.2. This error analysis goes beyond confusion matrices and κ scores: each unmatched event can be examined to determine the *types* of errors that caused the mismatches. This information helps determine which concrete changes to the annotation guidelines would be most effective.

Turning to the frame-based approach, the main purpose for our use-case is that—in combination with the event-based approach—it provides an indication of the nature of the disagreements between coders. In particular, if the frame-based approach yields higher agreement scores than the event-based approach, this suggests that the low agreement scores on the event-based approach are partly due to the following type of mismatches. Say C1 coded an entire sentence as 'down' on the *head y* tier, while C2 coded three separate long segments within that sentence as 'down', interspersed with two short 'neutral' segments. With the event-based method, all events coded on this tier would be regarded as unmatched. The frame-based method, on the other hand, would only count the disagreement of the 'neutral' segments; the rest would count as agreement.

The combination of the two approaches thus gives a more well-rounded overview of how the coders disagree. The event-based approach serves as a basis, supplemented by the frame-based approach. However, we should note that the frame-based approach, while in some cases providing an indirect indication of how coders disagreed, never provides a definitive insight into this important question.

Therefore, we propose to develop, in future work, an enriched version of the event-based method, which automatically categorizes the error-types of unmatched events (such as in the error analysis in Section 5.2 for the *head y* tier). This method would keep track of additional information such as the duration of the events that the coders agreed and disagreed on, and for each unmatched event, what type of error caused the mismatch. With this enriched event-based approach, the frame-based approach would become superfluous for our use-case, as the enriched event-based approach would provide all the necessary information to further improve the annotation procedure.

7. Conclusion

We evaluated guidelines for annotating NMMs by examining a test dataset involving two coders. We used a frame-based and an event-based approach to calculate inter-annotator agreement. Based on the results, we formulated concrete recommendations to further refine the annotation guidelines.

8. Acknowledgements

We thank Marc Schulder and James Trujillo for helpful discussion. We also gratefully acknowledge the three reviewers for their valuable feedback. This work is part of the project *Questions in Sign Language* (grant number VI.C.201.014, PI Roelofsen) financed by the Dutch Science Foundation (NWO).

9. Supplementary Materials

1. Annotation manual: https://doi.org/10.21942/uva.24080868
2. ELAN annotation template: https://doi.org/10.21942/uva.22732616
3. Inter-annotator agreement scripts: https://doi.org/10.21942/uva.24080724
4. Testset videos: https://doi.org/10.21942/uva.21666203
5. Testset annotation files: https://doi.org/10.21942/uva.22737074
6. Technical report: https://doi.org/10.21942/uva.25563540

10. Bibliographical References

Benjamin Bahan. 1996. *Non-manual realization of agreement in American Sign Language*. Ph.D. thesis, Boston University.

Roger Bakeman, Duncan McArthur, Vicenç Quera, and Byron F Robinson. 1997. Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2(4):357.

Roger Bakeman and Vicenç Quera. 2011. Sequential analysis and observational methods for the behavioral sciences.

Roger Bakeman, Vicenç Quera, and Augusto Gnisci. 2009. Observer agreement for timed-event sequential data: A comparison of time-based and event-based algorithms. *Behavior Research Methods*, 41(1):137–147.

Janet Bavelas and Nicole Chovil. 2018. Some pragmatic functions of conversational facial gestures. *Gesture*, 17:98–127.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

ELAN. 2023. Version 6.5. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. [link].

Lyke Esselink, Oomen Marloes, and Roelofsen Floris. 2023. Exploring new methods for measuring, analyzing, and visualizing facial expressions. *FEAST. Formal and Experimental Advances in Sign language Theory*, 5:35–48.

Aslı Göksel and Meltem Kelepir. 2013. The phonological and semantic bifurcation of the functions of an articulator: HEAD in questions in Turkish Sign Language. *Sign Language & Linguistics*, 16:1–30.

Santiago González-Fuente, Victoria Escandell-Vidal, and Pilar Prieto. 2015. Gestural codas pave the way to the understanding of verbal irony. *Journal of Pragmatics*, 90:26–47.

Henning Holle and Robert Rein. 2015. Easydiag: A tool for easy determination of interrater agreement. *Behavior Research Methods*, 47(3):837–847.

Andrea Lackner. 2017. *Functions of head and body movements in Austrian Sign Language*. De Gruyter Mouton, Berlin.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Carol Neidle. 2002. Signstream annotation: Conventions used for the American Sign Language linguistic research project. Technical report no. 11.

Naomi Nota, James P. Trujillo, and Judith Holler. 2021. Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11:1017.

Marloes Oomen, Lyke D. Esselink, Tobias de Ronde, and Floris Roelofsen. 2023. First steps towards a procedure for annotating non-manual markers in sign languages. In *NELS 53: Proceedings of the Fifty-Third Annual Meeting of the North East Linguistic Society*, volume 2, pages 257–266. GLSA.

Nina-Kristin Pendzich. 2020. *Lexical nonmanuals in German Sign Language. Empirical studies and theoretical implications*. De Gruyter Mouton.

Julius Sim and Chris C Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85(3):257–268.

Rosario Tomasello, Cora Kim, Felix R. Dreyer, Luigi Grisoni, and Friedemann Pulvermüller. 2019. Neurophysiological evidence for rapid processing of verbal and gestural information in understanding communicative actions. *Scientific Reports*, 9:16285.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.

Ronnie B. Wilbur. 2021. Non-manual markers – theoretical and experimental perspectives. In Josep Quer, Roland Pfau, and Annika Herrmann, editors, *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, pages 530–565.

Ronnie B. Wilbur and Cynthia G. Patschke. 1998. Body leans and the marking of contrast in American Sign Language. *Journal of Pragmatics*, 30:275–303.